2025-03-31

causalReg: Integrating Regression-based Causal Inference Tools for Empirical Research

Dylan Junluo CHEN
Department of Information Systems, College of Business, City University of Hong Kong

Signatories

Project team

- **Dylan Junluo CHEN** (Primary Investigator) is a PhD student in the Department of Information Systems at City University of Hong Kong, supervised by Prof. Ben Liu and Prof. Xin Li. His research focuses on open source community, data breaches and empirical methods. His homepage link is: https://chen-junluo.com/.
- **Bofei SHE** is a PhD student in the College of Computing at the National University of Singapore. With expertise in R programming, econometric methods, and statistical analysis, Bofei brings technical depth to the project.

Contributors

These contributors have provided extensive guidance to Dylan Junluo CHEN, and their expertise will be instrumental in helping him execute projects successfully:

- **Prof. Ben LIU** from the Department of Information Systems at City University of Hong Kong will provide econometric methodology guidance and validation.
- **Prof. Xin LI**, also from the Department of Information Systems at City University of Hong Kong, will contribute possible programming and package development consultation.

Consulted

- **Fellow PhD students** at the Department of Information Systems, City University of Hong Kong who use R as their primary analysis tool for empirical research and have expressed interest in contributing to the open source community.
- **Members of the China-R community** have been consulted regarding the needs of Chinese empirical researchers and the potential impact of this package.

Problem Definition

What the problem is, and Who it affects

R stands as a premier platform for empirical research due to its widespread adoption among economists and statisticians, robust statistical testing capabilities that support rigorous empirical work, and a thriving open-source community ideal for reproducible research. Despite these advantages, significant gaps persist:

For Chinese students and researchers:

1. Substantial language barriers limit access to predominantly English documentation
2. Absence of Chinese-language communities for support and knowledge sharing
3. Lack of integrated frameworks and educational materials in Chinese that connect theoretical concepts with practical implementation

For the international research community: While emerging empirical methods appear in various packages, they lack integration into a cohesive framework, forcing researchers to navigate multiple disconnected tools and creating inefficiencies in the empirical workflow.

Our proposal aims to establish an integrated regression-based causality development toolkit, alongside a dedicated Chinese-language support organization, community platform, and comprehensive educational materials. This approach addresses both technical and accessibility challenges in the empirical research workflow.

Have there been previous attempts to resolve the problem

As our Minimum Viable Product (MVP), we've selected one specific function for each of the three stages of causal inference:

1. **Omitted variable bias sensitivity analysis**(Oster 2019): The most comprehensive implementation is Stata's psacalc2 (https://github.com/ArthurHowardMorris/psacalc_supports_reghdfe). While it extends traditional psacalc with support for multiple fixed effects, it remains an unofficial patch that's cumbersome to use and lacks an R implementation.
2. **Heterogeneous treatment effect estimation**(Chernozhukov, Fernández-Val, and Luo 2018): The SortedEffects R package (https://cran.r-project.org/package=SortedEffects) currently leads in this area. However, it doesn't support continuous variables as predictors, severely limiting its applications. Additionally, the author appears to have left academia and doesn't respond to emails.
3. **Policy learning tools**(Athey and Wager 2021): While several implementations exist, such as policytree (https://github.com/grf-labs/policytree), they lack systematic Chinese-language tutorials and dissemination resources.

Why it should be tackled

This project directly aligns with ISC's mission to support open source infrastructure by:

1. **Expanding the R ecosystem's accessibility** - Creating bilingual tools and resources will significantly broaden R's reach within China's growing research community
2. **Strengthening research infrastructure** - An integrated empirical toolkit will provide essential infrastructure for producing reliable, reproducible research findings

By addressing these specific needs, our project will strengthen both the R ecosystem and the broader open source research infrastructure, creating lasting impact on empirical research practices in China and beyond.

The proposal

Overview

The causalReg package will provide an integrated toolkit for regression-based causal inference methods in R, specifically designed to facilitate adoption by China's empirical research community. By addressing the methodological gaps identified in the problem statement, causalReg will:

1. **Integrate critical causal inference techniques** into a cohesive framework built on established R packages like lfe (Gaure 2013)
2. **Provide bilingual documentation and tutorials** targeting Chinese research students
3. **Establish ongoing maintenance** for key capabilities including SortedEffects
4. **Create a bridge** between Stata-centric workflows and R's powerful ecosystem

This project directly aligns with the R Consortium's mission by expanding R's reach in a significant academic community, ensuring package sustainability, and promoting reproducible research practices.

Detail

The causalReg package will be developed in three progressive phases, each targeting a key component of the regression-based causal inference workflow:

Phase 1: Omitted Variable Bias Sensitivity Analysis

Building on the methodology developed by Oster (2019), we will implement a robust solution for sensitivity analysis that:

- Extends existing implementations to support multi-way fixed effects through integration with the lfe package
- Provides intuitive visualization of sensitivity bounds
- Includes comprehensive documentation with applied examples from economics and information systems

The implementation will handle both continuous and binary outcomes, with full support for the standard econometric workflow including clustered standard errors and various fixed effects specifications.

Phase 2: Heterogeneous Treatment Effect Estimation

We will revive and enhance the SortedEffects package (Chernozhukov, Fernández-Val, and Luo 2018) to:

- Fix critical bugs affecting continuous variable analysis
- Improve integration with standard regression workflows
- Add robust visualization capabilities for heterogeneous effects
- Enhance documentation with practical examples

This module will allow researchers to move beyond average treatment effects to understand how causal impacts vary across the population - a critical capability for policy analysis.

Phase 3: Policy Learning Tools

Building on recent methodological developments in machine learning for causal inference (Athey and Wager 2021), we will:

- Implement methods for optimal treatment targeting
- Provide tools to estimate heterogeneous treatment effects with machine learning
- Create accessible interfaces for policy learning without requiring deep ML expertise
- Develop step-by-step tutorials for applied policy analysis

Minimum Viable Product

The MVP for causalReg will include:

1. A functioning R package with robust implementations of Oster sensitivity analysis and SortedEffects
2. Comprehensive bilingual documentation including:
   - Function references
   - Methodological vignettes
   - Applied tutorials with real-world datasets
3. Integration with existing R econometric workflows
4. Initial educational materials for Chinese users

Architecture

The package will be built with a modular design that:

1. Leverages established packages (lfe, sandwich, ggplot2) for core functionality
2. Provides consistent syntax across different causal inference methods
3. Includes flexible visualization components
4. Offers both high-level interface functions and access to underlying components

Project plan

Start-up phase (Month 1)

Set up GitHub repository with MIT license, contribution guidelines, and continuous integration pipeline; establish bilingual documentation standards; create project website structure.

Technical delivery

Phase 1: Omitted Variable Bias Sensitivity Analysis (Month 2-3)

- **Month 2:** Implement Oster methodology with fixed effects support and visualization components.
- **Month 3:** Complete bilingual documentation with vignettes and submit initial release (0.1.0) to CRAN.

Phase 2: Treatment Effect Heterogeneity Analysis (Month 4-5)

- **Month 4:** Fix SortedEffects package's continuous variable handling and enhance visualization capabilities.
- **Month 5:** Integrate with causalReg package and develop bilingual tutorial with applied examples.

Phase 3: Policy Learning Implementation (Month 6-7)

- **Month 6:** Implement core policy learning algorithms with interface for optimal treatment targeting.
- **Month 7:** Complete package integration with comprehensive documentation and release version 1.0.0.

Other aspects

Community Building and Promotion (Ongoing)

- **Month 3:** Launch announcement blog post and WeChat campaign.
- **Month 5:** Host bilingual webinar demonstrating first two modules.
- **Month 7:** Release comprehensive video tutorial series.
- **Month 8:** Develop university workshop materials and training sessions.

Reporting and Communication: Monthly progress reports to ISC, quarterly blog posts on R Consortium website, and final project report with case studies.

Requirements

People

Core Team

The project requires a small, focused team with expertise in both econometrics and R programming:

- **Dylan Junluo CHEN** (PhD student, City University of Hong Kong, Department of Information Systems): Project lead with experience in empirical research methods and R package development
- **Bofei SHE** (PhD student, National University of Singapore, College of Computing): Programming expertise and econometric knowledge

Additional support will come from two professors and PhD colleagues at CityU with experience in R programming and empirical research who have expressed interest in contributing to open-source projects.

Processes

The project will implement GitHub-based development with branch protection, continuous integration testing, and documentation requirements for all new features. Communication will occur through bi-weekly team meetings, with monthly progress updates to ISC.

Tools & Tech

The project will utilize established R packages (lfe, ggplot2, sandwich, glmnet, tidyverse) as dependencies, with GitHub for code hosting, CRAN for distribution, and GitHub Pages for documentation.

Funding

The requested funding of $1,000 will be allocated as follows:

1. **Educational Materials ($300)**:
   - Video tutorial production
   - Example dataset preparation
   - Translation services for documentation
2. **Community Building ($300)**:
   - Virtual workshop hosting
   - WeChat official account maintenance
   - User community support
3. **Contingency ($400)**:
   - Reserved for unexpected requirements

Summary

This project requires a focused development team with expertise in both econometrics and R programming, supported by academic advisors with domain knowledge. The technical requirements are modest, primarily relying on existing R infrastructure and packages. The funding requested will primarily support dedicated development time to ensure timely completion of all components and the creation of high-quality educational materials necessary for adoption by the target user community.

Success

Definition of done

The causalReg project will be considered complete when:

1. A fully functional R package is published on CRAN containing all three modules (omitted variable bias sensitivity analysis, treatment effect heterogeneity analysis, and policy learning implementation) with comprehensive test coverage.
2. Bilingual documentation is available with function references, methodological vignettes, applied tutorials, and video demonstrations accessible to both English and Chinese users.
3. A sustainable community structure is established through an active WeChat group, responsive GitHub issue process, and clear contribution guidelines.
4. Educational materials are available including workshop materials, self-paced resources, and example datasets with complete workflows.

Measuring success

We will track project success through:

Short-term metrics (1 year):

- Technical adoption: 100+ GitHub stars, 1,000+ CRAN downloads, 90% bug resolution within 2 weeks
- Community engagement: 100+ WeChat group members, 5+ articles with 1,000+ views, 5+ tutorials with 500+ views

Medium-term metrics (2 years):

- Academic impact: 5+ papers citing the package, 3+ university courses using materials, 2+ external contributions
- Sustained growth: 5,000+ cumulative downloads, active issue discussions, ongoing feature development

Qualitative indicators: Positive feedback from Chinese researchers, reduced barriers to R adoption, and examples of research utilizing previously inaccessible methods.

Future work

This project lays the foundation for several future extensions:

- **Methodological expansion**: Implement synthetic control methods, instrumental variable approaches, regression discontinuity designs, and machine learning-based causal inference techniques as additional modules.
- **Educational resources**: Create a full online course, textbook companion materials, and university curriculum integration guides to support broader adoption.
- **Community development**: Establish annual contributor workshops, regional user groups in major Chinese cities, and regular webinar series featuring applied case studies.

These extensions will build upon the core infrastructure established in the initial project, leveraging the bilingual framework and expanding the toolkit's capabilities while maintaining the focus on accessibility for Chinese researchers.

Key risks

People

- **Risk**: Minimal personnel risk as project aligns with team members' thesis topics and career aspirations, ensuring sustained motivation and commitment throughout development.

Processes

- **Risk**: Bilingual documentation synchronization challenges.
- **Mitigation**: Establish streamlined translation workflow with prioritized documentation elements.

Tooling & Technology

- **Risk**: Dependency on third-party packages for core functionality.
- **Mitigation**: Monitor upstream changes and maintain contingency forks if necessary.

Costs

- **Risk**: Unexpected costs for community outreach.
- **Mitigation**: Leverage existing academic networks and allocate contingency funds.

Athey, Susan, and Stefan Wager. 2021. "Policy Learning with Observational Data." *Econometrica* 89 (1): 133–61.

Chernozhukov, Victor, Iván Fernández-Val, and Ye Luo. 2018. "Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages." *Econometrica* 86 (6): 1911–38.

Gaure, Simen. 2013. "Lfe: Linear Group Fixed Effects."

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204.